[連載]フリーソフトによるデータ解析・マイニング　**第11回**

Rと仮説検定

1. 仮説検定

1.1 仮説検定とは

通常、データ解析に用いるデータは標本データであるが、研究対象は母集団である。よって、データ解析では標本データの統計量(比率、平均、分散など)を用いて母集団の特性・母数(比率、平均、分散など)を説明する。しかし、標本の数、標本の偏りなどの影響で標本データの統計量が必ずしもそれに対応する母数を正しく説明できるとは限らない。

仮説検定は、比較する母数には差がないという仮説を立てておき、その仮説の是非を確率理論に基づいて判定を行うデータ解析方法である。仮説検定では常に比較するものには差がないという仮説を立てるので、仮説検定を帰無仮説検定と呼ぶが、本稿では略して仮説検定と呼ぶ。仮説が正しいと判定されたときには**仮説が採択**され、仮説が正しくないと判定された場合は**仮説が棄却**されると言う。

1.2 片側検定と両側検定

仮説検定では、まず仮説と対立仮説を決めなければならない。例えば、標本の母数と既知の母数が同じであると見なすことができるかどうかに関する問題では次のような仮説を立てる。

H_0：　既知の母数=標本の母数

通常、仮説は記号 H (Hypothesis,仮説)を用いる。上記の仮説を否定する仮説としては

H_1：(a) 既知の母数≠標本の母数
　　(b) 既知の母数<標本の母数
　　(c) 既知の母数>標本の母数

がある。この H_1 を仮説 H_0 の**対立仮説**と言う。

対立仮説(a)に関する検定を両側検定、仮説(b)、(c)のいずれか１つの対立仮説に関する検定を片側検定と言う。2標本A、Bの母数の差の検定に関する仮説の形式を次に示す。

H_0：　標本Aの母数=標本Bの母数

1.3　2 種類の過誤と有意水準

仮説検定は、標本データから母集団に対して仮説を立証するので 100%正しい結論はありえなく、誤りを犯す可能性は十分あり得る。

仮説が違わないのに違ったと判断した誤りを**第Ⅰ種の過誤**、違っているのに違わないと判定された誤りを**第Ⅱ種の過誤**と言う。通常の仮説検定では第Ⅰ種の過誤のみをチェックする。この第Ⅰ種の過誤の確率を**有意水準**と言い、αで表す。有意水準がαのとき「1 - α」を**信頼係数**と呼ぶ。有意水準αの値は 0.01(1%)、0.05(5%)、0.1(10%)が多く用いられている。

有意水準 0.05(5%)で仮説検定を行った場合、仮説が採択(仮説がもし成立する)されるならば、100 回の中「違わないのに違ったと判断する」誤りを 5 回犯す可能性がある。有意水準が大きければ大きいほど第Ⅰ種の過誤を犯す可能性、つまり仮説が間違って棄却される可能性が高くなる。有意水準を小さくすると第Ⅰ種の過誤を犯す確率は減少するが、そのかわりに第Ⅱ種の過誤を犯す確率が増加する。

対立仮説に対応する両側検定、片側検定の採択域、棄却域及び有意水準αについて標準正規分布を用いたイメージを図 1、2、3 に示す。

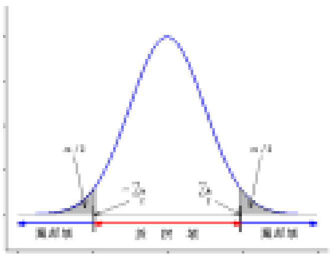

図 1　両側検定の採択域

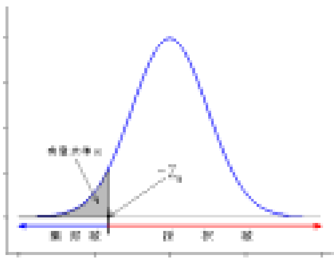

図 2　左片側検定の採択域

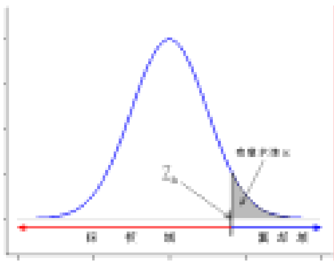

図 3　右片側検定の採択域

1.4 検定統計量と *p* 値

仮説検定では、検定を行う統計量によって計算式が異なる。説明の便利のため標本平均を考えよう。

区間推定を説明するときに、標本 $X_1, X_2, \cdots X_n$ がお互いに独立で正規分布 $N(\mu, \sigma^2)$ に従う確率変数の平均 $\overline{X}$ は正規分布 $N(\mu, \frac{\sigma^2}{n})$ に従い、次のように変換された Z は標準正規分布に従うことについて既に説明した。

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

また、標本サイズが小さいとき母集団の標準偏差のかわりに標本の標準偏差を用いると自由度 $n-1$ の t 分布に従う。

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

仮説検定はこのような関係式を用いてその確率値を求めて、仮説について判定を行う。

例えば、全国の 17 際男子の平均身長は 170.7cm であるとする。某高校の 17 際男子生徒 25 人の身長を測ったところその平均が 172.58cm、分散が 34.27 であったとする。身長のデータが正規分布に従うと仮定した場合、次のT は自由度$n-1$のt分布に従う。

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{172.58 - 170.7}{\sqrt{34.27}/\sqrt{25}} \cong 1.6057$$

この値をt値、あるいはt検定統計量と呼ぶ。確率変数がF 分布に従う際には、F 値あるいはF 検定統計量と呼ぶ。

検定統計量に対応する上側確率、あるいは下側確率を p **値**と呼ぶ。Rでは上記の例の両側検定におけるt値 1.6057 に対応する p 値は次のように求める。

```
> 2*pt(1.6057,24,lower.tail=F)
[1] 0.121419
```

1.5 仮説検定のプロセス

1.5.1 仮説と対立仮説を決める

仮説検定では、常に比較するものには差がないという仮説を立てる。前項の身長の例では、「某高校の 17 際男子生徒の身長の母平均は全国の 17 際男子の平均身長 170.7cm(μ)と同じである」が仮説となる。つまり、仮説H_0 は「μ＝標本の母平均」である。

検定を行う際には、まず対立仮説を決めなければならない。もし、上記の例で「某高校の 17 際男子生徒の平均身長が全国の 17 際男子の平均身長と同じであると言えるであろうか？」を考える場合は、ともかく異なることを立証すればよいので、対立仮説は「μ ≠ 標本の母平均」を取り、両側検定を行えばよい。

1.5.2 有意水準を決める

有意水準は恣意的な値で、業界と問題によって異なる。通常では 0.01(1%)、0.05(5%)、0.1(10%)が多く用いられている。

1.5.3 検定統計量と p 値を求める

データが従う分布や行う検定などを吟味した上で検定統計量及び p 値を求める。前項の例で

は、 *t* 値は 1.6057 で、 *p* 値は 0.1214 である。

1.5.4　仮説の採択・棄却を検討する

仮説検定は、検定統計量が有意水準αの下で採択域に入ると仮説が採択され、棄却域に入ると仮説が棄却されると判定する。

前項の例について、有意水準αを 0.05 として検定を行うことにする。自由度(標本数―1)が 24 である *t* 分布の両側の採択域・棄却域を図 4 に示す。図 4 でわかるように求めた *t* 値 1.6057 が採択域に入っているので仮説が採択されると判断する。つまり、有意水準 0.05 ではこの高校の 17 際男子生徒の身長が 17 際男子の全国平均と異なるとは言えないと判断する。

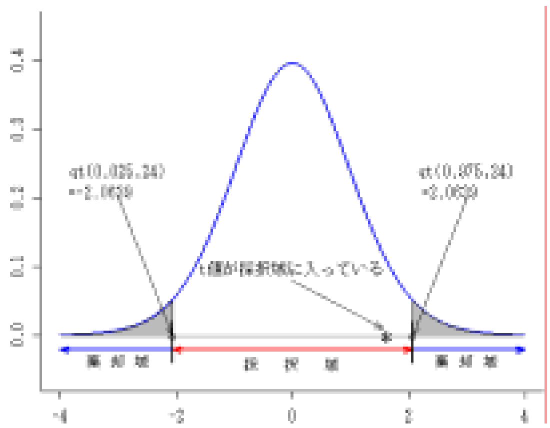

図 4　*t*(0.05,24) 分布の両側検定の採択域

p 値と有意水準αを用いて仮説の採択・棄却を判断することも可能である。「*p* 値<有意水準α」であると仮説が棄却される。例えば、上記の例では「*p* 値>有意水準」(0.1214 >0.05)であるので有意水準 0.05 で仮説が採択される。

データ解析ソフトでは、検定の種類や条件を指定すると自動的に検定統計量や *p* 値など検定に必要となる情報が返されるようになっている。

2．Rにおける仮説検定関数

Rには古典的検定(Classical Tests) パッケージ ctest があり、その中には 20 数種類の検定関数が用意されている。パッケージ ctest はRをインストール際に自動的にインストールされるようになっている。検定の関数と書式のマニュアルは、Rのメニュー⇒「Help」⇒「Html help」⇒「Packages」⇒「ctest」から辿りつくことができる。

2.1　1 標本の検定

1 標本の検定とは、1 つの標本データと 1 つの母数が既知であり、標本が既知の母数の母集

団に属するかどうかに関する検定である。本項では、平均、比率の検定のみを取り上げる。

2.1.1　母平均の検定

Ｒのパッケージ ctest には平均に関する *t* 検定を行う関数 t.test が用意されている。関数 t.test は前節で説明した母平均の *t* 検定を行うことが可能である。関数 t.test の書式と引数を次に示す。表 2 に関数 t.test に用いる引数の設定とデフォルトの設定を示す。

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE,
var.equal = FALSE, conf.level = 0.95, ...)
```

表 2　t.test の引数

主な引数	引数の設定	デフォルト
x, y(データ)	データを代入	y=NULL
var.equal(等分散)	TRUE、FALSE	FALSE
Paired(対応性)	TRUE、FALSE	FALSE
Alternative(対立仮説)	two.sided、less、 greater	two.sided
mu(母平均)	既知の母平均	0
conf.level(信頼係数)	自由	0.95

1 標本の母平均の検定では、paired 、var.equal はデフォルトの設定を変える必要はない。次に例を用いてその使用方法を示す。

平成 15 年度文部科学省の調査によると全国の 17 際男子の平均身長は 170.7cm である。某高校 17 際の男子 25 人について調査を行ったところ次に示すデータが得られたとする。ここでは関数 t.test と標本データを用いて前節と同じの検定を行うことにする。データの入力と関数 t.test の使用例を次に示す。

```
>x<-c(171.6, 173.6, 167.6, 169.1, 183.0, 173.7, 168.3, 169.9, 182.3, 166.0, 172.8, 184.1, 158.9,
168.1, 168.5, 175.3, 179.6, 170.7, 173.1, 173.6, 169.1, 167.9, 177.8, 171.8, 178.1)
>t.test(x,mu=170.7,alternative ="two.sided")
```

```
        One Sample t-test
data:   x
t =1.6057 , df = 24, p-value =0.1214
alternative hypothesis: true mean is not equal to 170.7
95 percent confidence interval:
 170.1635 174.9965
sample estimates:
mean of x
   172.58
```

返された結果の *t* 値は 1.6057 で、 *p* 値(p-value)は 0.1214 である。これは、前節で取り上げた例の結果と一致している。前節の例ではこの標本データの平均と分散を用いていることに気づいて欲しい。 *p* 値が有意水準 0.05 より大きいので仮説「既知の母数(170.7)=標本の母平均」

が採択される。

関数 t.test では、信頼区間(95 percent confidence interval)をも返す。両側の 95%の信頼区間は[170.1635, 174.9965]である。

2.1.2　母比率の検定

母比率に関する検定としては関数 binom.test と prop.test が多く利用されている。次に関数 prop.test の書式を示す。

```
prop.test(x, n, p = NULL,alternative =c("two.sided", "less", "greater"),conf.level = 0.95, correct = TRUE)
```

引数alternative を略した場合は両側検定､conf.level を略した場合は有意水準0.05 (5%)の結果を返す。

例えば、現政権の支持率に関する調査では、現政権の支持率は 45%であったが、新たの調査では 1000 人の中で 470 人が現政権を支持するという結果が得られたとする。2 回の調査の現政権に対する支持率が同じであると言えるであろうか？

この比率の検定は関数 prop.test を次のように用いて行うことが可能である。

```
> prop.test(470,1000,p=0.45)
        1-sample proportions test with continuity correction
data:   470 out of 1000, null probability 0.45
X-squared = 1.5364, df = 1, p-value = 0.2152
alternative hypothesis: true p is not equal to 0.45
95 percent confidence interval:
 0.4387437 0.5014896
sample estimates:
   p
0.47
```

返された p 値 0.2152 は有意水準 0.05 より大きいため、有意水準 0.05 では仮説が採択される。

2.2　2 標本の場合

2.2.1　平均の差の検定

2 標本の平均値の差の検定は関数 t.test を用いて行うことができる。表 1 からわかるように平均の差の検定では var.equal(等分散)について設定を行わなければならない｡パッケージ ctest には 2 標本の等分散に関して検定を行う関数 var.test が用意されている。

```
var.test(x, y, ratio = 1, alternative =c("two.sided", "less", "greater"), conf.level = 0.95, ...)
```

引数 x,y は 2 つの標本データベクトル、ratio = 1 は標本 x,y の母分散の比で、alternative は対立仮説を指定する引数である。次に簡単なデータセットを入力し、その使用形式を示す。

```
>x<-c(1,2,3,4,5)
>y<-c(1,5,9,13,17)
>var.test(x,y)
        F test to compare two variances
data:  x and y
F = 0.0625, num df = 4, denom df = 4, p-value = 0.01995
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.006507346 0.600283118
sample estimates:
ratio of variances
            0.0625
```

p 値が 0.05 より小さいので、有意水準 5%では、「母分散が同じである」という仮説が棄却される。

検定関数 t.test、var.test などでは、次に示すデータセット sleep のようなデータフレーム形式をも扱うことができる。データ sleep はそれぞれ１０人からなる 2 つのグループに対し、異なる睡眠剤の効果を調べたデータである。

```
> data(sleep)
> sleep
   extra group
1    0.7     1
2   -1.6     1
＜中略＞
19   4.6     2
20   3.4     2
```

第 1 列はデータの番号で、第 2 列「extra」は睡眠時間の増加で、第 3 列「group」は試験を行う際に分けたグループの番号である。

この両グループの平均には差があると言えるかどうかについて有意水準 5%で仮説検定を行ってみよう。平均の差について検定を行う前に、まず 2 標本の等分散について検定を行うことにする。

```
> var.test(extra~group,data=sleep)
        F test to compare two variances
data:   extra by group
F = 0.7983, num df = 9, denom df = 9, p-value = 0.7427
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
 0.198297 3.214123
sample estimates:
ratio of variances
         0.7983426
```

返された結果の p 値は 0.7427 で有意水準 0.05 より大きいので、有意水準 0.05 で「両グループの母分散は同じである」という仮説が採択される。この結果は、2 標本の平均の差の検定を行う際に等分散の設定を「var.equal = TRUE」にする根拠となる。次に 2 標本の平均の検定の計算結果を示す。

```
>t.test(extra~group,var.equal = TRUE,data=sleep)
        Two Sample t-test
data:   extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3638740   0.2038740
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

返された結果の p 値 0.07919 は有意水準 0.05 より大きいので、「両グループの睡眠時間の平均値には差がない」という仮説が有意水準 0.05 で採択される。しかし、有意水準が 8%の場合は p 値 0.07919 が 0.08 より小さいので仮説が棄却される。

この検定では 2 標本の母集団が正規分布に従っていることを仮定している。しかし、実際の問題では分布の正規性に関する前提条件が不十分である場合もある。分布の仮定を置かずに行う検定としては、パッケージ ctest の中のウィルコクソン(Wilcoxon)検定を行う関数 wilcox.test が多く用いられている。wilcox 検定は Mann-Whitney の U 検定と同じく順位検定を行うが、標本の数をやや多くとると t 検定と同等の検出力を持つと言われている。

2.2.2　比率の差の検定

2 標本の比率検定は関数 prop.test を用いて行うことが可能である。例えば、現政権の支持率に関する 2 回の調査で、第 1 回では 1000 人の中 450 人が、第 2 回では 1000 人の中 470 人が支持するデータが得られたとする。有意水準 0.05 での比率の差の両側検定は次のように prop.test を用いることができる。

```
>prop.test(c(450,470),c(1000,1000))
        2-sample test for equality of proportions with continuity correction
```

```
data:  c(450, 470) out of c(1000, 1000)
X-squared = 0.7267, df = 1, p-value = 0.394
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06467686  0.02467686
sample estimates:
prop 1 prop 2
  0.45   0.47
```

2.3 カイ 2 乗検定

定性的データ(カテゴリカルデータ)の解析を行う際にはしばしば度数データを用いる。その例として、本連載第 4 回(2003 年 11 月号)で用いた調査データの例の一部を再掲する。次の質問について異なる 3 つのグループに対して行った調査の集計結果を表 2 に示す。

質問：　今の生活がむなしく感じられることがある。
1.　よくある　　2.　ときどきある　3.　あまりない　4.　ない

表 2　質問の回答結果(合計 1021 人)

グループ	回答 1	回答 2	回答 3	回答 4	合計
A	101	120	70	35	326
B	153	162	88	46	369
C	89	135	78	24	326

このような調査データを解析する際には「今の生活がむなしく感じられる」という傾向がグループごとに異なるかに関して関心を持つ。その解析方法の一つとして χ^2 (Chi-squared, カイ 2 乗)検定が広く知られている。パッケージ ctest にはカイ 2 乗検定の関数 chisq.test が用意されている。データセットが表 2 のような分割表形式であればデータセットを関数に投入するだけで計算結果が返される。次に表 2 のデータセットの作成、chisq.test の使用形式及び返された結果を示す。

```
>muna<-matrix(c(101,120,70,35,153,162,88,46,89,135,78,24),3,4,byrow=T)
> chisq.test(muna)
```

```
        Pearson's Chi-squared test
data:   muna
X-squared = 8.338, df = 6, p-value = 0.2144
```

返された p 値 0.2144 は通常広く使用されている有意水準 0.05、0.1 より大きいので、有意水準 0.05 で「むなしく感じる傾向が同じである」という仮説が採択される。

3. 仮説検定における注意点と参考資料

仮説検定における p 値は、サンプル数に依存する。サンプル数を増やせば幾らでも p 値を小さくできる。また、恣意的な有意水準はどこまで意味を持っているかも一つの問題である。例えば、有意水準を 0.05 とした場合、 p 値が 0.051 と 0.049 はわずか 0.002 の差であるが完全に異なる結果になってしまう。

仮説検定のアプローチでデータ解析を行う際には誤った使い方がないように十分気を付けなければならない。

最近Rを用いた統計やデータ解析に関するインターネット上の資料の公開や和書が続々刊行・企画され、Rが急速に普及しつつあることが実感されている。主なホームページアドレスと和書を次にリストアップする。

- 群馬大学社会情報学部青木繁伸教授の次のホームページには豊富でかつ信頼性の高い統計に関する資料やＲのプログラムが公開されている。http://aoki2.si.gunma-u.ac.jp/R/
-
- Ｒによる統計解析の基礎、 中澤港（著)、ピアソンエデュケーション。
- 工学のためのデータサイエンス入門 －フリーな統計環境 Ｒ を用いたデータ解析、間瀬茂・神保雅一・鎌倉稔成・金藤浩司(共著)、数理工学社。
- The R Bookーデータ解析環境Ｒの活用事例集ー、岡田昌史・中間栄治・その他(共著)、九天社。
- フレッシュマンから大学院生までのデータ解析・Ｒ言語、渡辺利夫(著)、ナカニシヤ出版（９月頃発行予定？）